DOCUMENT RESUME

ED 279 680 TM 870 067

AUTHOR Haney, Walt; Madaus, George
TITLE Effects of Standardized Testing and the Future of the National Assessment of Educational Progress. Working Paper for the NAEP Study Group.
PUB DATE Sep 86
NOTE 37p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Educational Assessment; Educational Policy; Educational Testing; *Educational Trends; Elementary Secondary Education; Evaluation Utilization; *Futures (of Society); Higher Education; National Surveys; Standardized Tests; *Testing Problems; Testing Programs; *Test Use
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

Topics related to the National Assessment of Educational Progress (NAEP) are discussed. Issues affecting the impact of testing on education include: what is tested; how scores are referenced; internal versus external sources of testing; and the rewards, sanctions, or stakes associated with test results. Testing has both intentional (direct) and indirect effects. Seven principles regarding the impact of testing are especially true when the tests are used for important social decisions: (1) measuring any social indicator results in its distortion or change in status; (2) the power of testing is determined by the perception of its effects; (3) teachers will teach to the test; (4) previous tests influence curriculum content; (5) instruction and learning are adjusted to test format; (6) test results become a major goal of education; and (7) control of the curriculum is transferred to the testing agency. NAEP can use test information: (1) to inform policy makers about the current state of education, or (2) as administrative devices in policy implementation. NAEP is the most valid source of national achievement data and has been designed to inform educational policy making, rather than to direct it. Increasing its policy-making ability might have distorting effects or threaten test validity. It has been, and will continue to be useful as a model for other testing activities. (A list of ERIC references concerning assessment is included.) (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279680

Effects of Standardized Testing and the Future of the National Assessment of Educational Progress

Working Paper for the NAEP Study Group

by

Walt Haney and George Madaus
Center for the Study of Testing, Evaluation and Educational Policy
Boston College
Chestnut Hill, MA 02167

ph 617/552-4521

September 1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Haney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 067

Paper commissioned by
THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT
1986

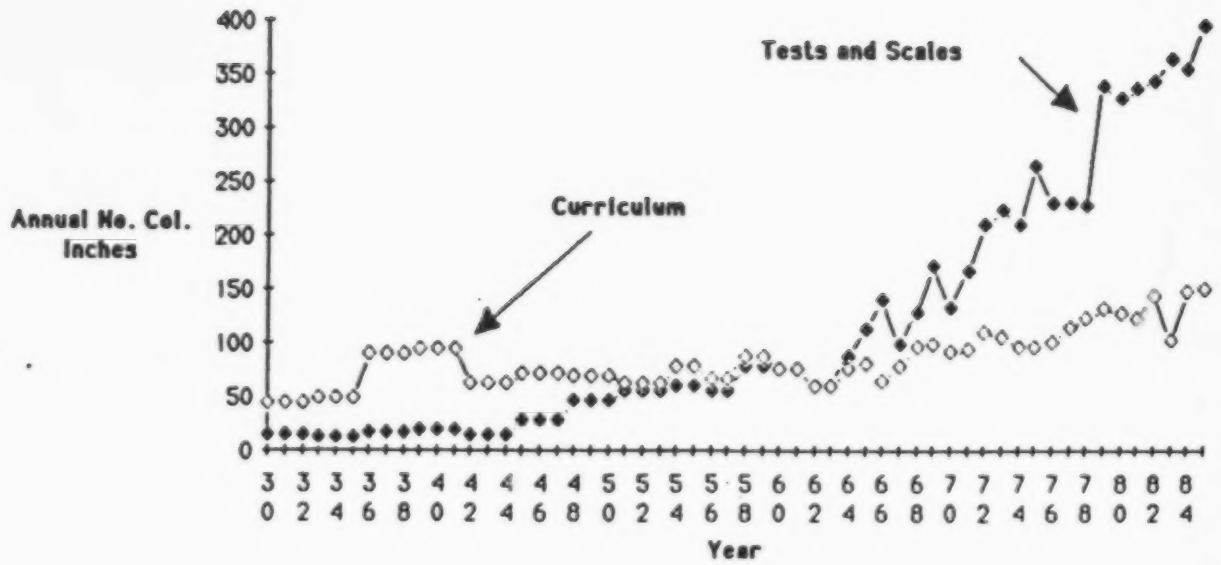
I. Introduction

> Since tests have a powerful directive influence on teaching and the study of pupils, a major policy to follow is to establish a testing program that faithfully reflects the objectives sought by the school. In this way the influence of testing is to reinforce the objectives sought by the school.
>
> Ralph Tyler, "What Testing Does to Teachers and Students," 1959, In Anastasi, 1966, p. 49.

Standardized testing in the schools is on the increase. As Pipho (1985, p.19) observed "nearly every large education reform effort of the past few years has either mandated a new form of testing or expanded uses of existing tests." The increasing prominence of testing over the last five years is linked directly to efforts to reform education, particularly at the state level. For example, a 50-state survey of reform measures conducted by Education Week found that: 29 states required competency tests for students, and 10 other states had such a requirement under consideration; 15 states required an exit test for graduation, 4 additional states had such a measure under consideration; 8 states employed a promotional "gates" test, while 3 others were considering such a mandate; finally, 37 states had some sort of state assessment program, and 6 additional states had such a program under consideration. This growing use of tests in the policy sphere by agencies external to the local education agencies (LEAs), we will argue, is having increasing impact on what is taught and learned in schools.

As a means of documenting this increasing attention to testing, and contrasting it with curriculum concerns, we charted the amount of space in Education Index over the last 50 years devoted to citations concerning testing and curriculum. As shown in Figure 1, the average annual number column inches devoted to citations concerning curriculum has increased only modestly over the last half-century -- from 50 - 100 inches per year in the

Figure 1: Education Index Listings Under Testing and Curriculum 1930 - 1985



1930s and 40s to only 100 - 150 in recent years. In contrast, attention devoted to testing in Education Index has increased dramatically, from only 10 - 30 column inches in the 1930s and 1940s to well over 300 inches in the 1980s.

Measuring column inches in Education Index is, of course, a fairly crude way of charting what is happening in the world of education, but these data certainly suggest a trend that we suspected even before looking at the Education Index. It is that increasingly, standardized testing seems to have become the coin of the educational realm. In recent years, it seems that the aims of education and the business of our schools are addressed not so much in terms of curriculum - the courses of study that are or should be followed -- as in terms of what gets tested. The data from the Education Index showing that the relative attention to curriculum and testing issues has undergone a ten-fold change in the last 50 years, clearly suggests this.

Before reviewing arguments and evidence bearing on the impact of testing, we need to comment briefly on what is meant by the term "standardized tests." Essentially by this term we refer to standardized achievement tests of reading and math skills, including state- and local-education-agency sponsored standardized tests, commercially developed nationally normed tests, and tests that have sometimes been called basic skills or minimum competency tests. Under the rubric of "standardized achievement tests" one might even include the Scholastic Aptitude Test (SAT) which is, we think, essentially a general achievement test of verbal and math skills rather than a test of aptitude (though we will not in this brief presentation attempt to delve into the controversy over the aptitude/achievement distinction, either in general or with respect to the SAT).

By way of introduction it should be noted that such tests may be used for a variety of purposes, including general evaluation of schools and programs, diagnosis of student strengths and weaknesses, student grade promotion, high school graduation, guidance, college admissions and even teacher evaluation. However, despite the initial purposes of either test developers or sponsors for achievement testing, , once a test is administered the results often get used for other quite different purposes. Perhaps the most prominent example of this is the way in which college admissions tests, despite occasional protestations by test sponsors, have come to be used as general indicators of the educational health of states, and the nation as a whole (in this regard see, for example, the United States Office of Education's annual Wall Chart, or the recent study, Trends in Achievement, by the Congressional Budget Office of the U.S. Congress, 1986. Both use SAT and ACT test data as two sources of evidence regarding national trends in achievement).

Thus, in this brief paper we will not attempt to distinguish between the different, ostensible, commonly recognized, purposes of standardized tests. Instead we will attempt to set out some fairly broad ideas about the effects of testing on education, and on curriculum in particular; offering some ideas on the past and future of the National Assessment of Educational Progress (NAEP). Specifically the following four sections of this paper are organized around these topics:

II. Conditions of Testing Affecting Its Impact

III. Seven Principles Regarding the Impact of Testing

IV. NAEP as an Instrument for Informing Educational Policy; and

V. The Future of NAEP

II. Conditions of Testing Affecting Its Impact

Rather than commenting on the myriad specific ostensible uses of testing, instead, in this section, we offer some general observations on four broad issue that condition the impact of testing: (1) What is tested, (2) How scores are referenced; (3) The source of testing; and (4) The rewards or sanctions associated with test results

2.1 What is tested

A wide range of variables has been the subject of measurement, though the main emphasis has been on the measurement of cognitive rather than affective characteristics. The cognitive variables which have attracted most attention are "intelligence", and achievement in basic skill areas of the curriculum (reading , arithmetic). As one proceeds up the educational ladder into secondary schools, where instruction is organized around subject matter/content areas, rather than around specific skills, commercially available test batteries become less specific and less related to what is taught. High school test batteries closely resemble elementary school batteries in that they are more oriented to the basic skills of numeracy and literacy than to what is taught in specific subject fields like math, physics, history, English literature, etc. As a result, such test scores are less relevant to the work of the high school teacher. However, some of the recent reform reports call for the development of exams for specific secondary school curriculua areas. This has profound implications for curriculum and instruction at that level. This is because in general, it seems that other

things being equal, the impact of testing is greater when tests are keyed to specific courses.

2. How scores are referenced.

The main distinction here is between norm-referenced tests, on which performance is assessed by reference to the performance of other students, and criterion-referenced tests, on which performance is assessed by reference to the mastery of specific content domains. It should be noted that norm-referenced information can also be structured to provide criterion-referenced interpretations and vice versa. While criterion-referenced tests are increasingly hailed as superior to norm-referenced ones in terms of information provided to teachers, norm-referenced information is valuable for comparative purposes. Further, the specificity of criterion-referenced information from commercially available tests, relative to what is actually taught at the local level can often be dubious.

The general point to be made here is that however tests are referenced, if they are poorly matched to what is taught in schools, and if they are linked to important decisions, they can have great impact, for good or ill.

3. Internal vs. external testing programs.

An important factor regarding the impact of testing is its **source**. The main distinction here is between internal testing, and external testing. An internal testing program is one which is carried out within a school at the initiative and under the control of the school superintendent, principal or teacher. In this category we include the traditional norm-referenced standardized achievement testing programs that have been used by school

systems since the 1920's, as well as the use of the newer commercially available criterion referenced achievement tests. Internal testing is not limited to off-the-self commercial tests; it also includes traditional teacher made tests which we will not treat in this paper. In larger systems tests can be built by school system personnel or customized by a contractor for the system.

External testing programs are those controlled and/or mandated by an external authority, such as a state education education agency (SEA), the state legislature or a private agency such as the College Entrance Examination Board (CEEB). It can be a semi-voluntary test like the Scholastic Aptitude Test (SAT). More commonly, it is a test which the state mandates if a student, teacher or school district are to fulfill certain mandated requirements. State minimum competency exams linked to graduation, promotion or remediation decisions, the English 'O' and 'A' Levels, the Irish Leaving Certificate Examinations, and the New York State Regents Exams are all examples of the latter. External testing programs can use norm- or criterion-referenced tests; use off-the-shelf commercially available instruments, tests built by the SEA, or those developed for such a sponsor by a contractor.

External tests have long been a part of European educational systems. Until the advent of minimum competency testing, most educational testing in the United States had been internal in origin and control. External testing programs were relatively rare; the New York Regents Exams and the CEEB examinations being notable exceptions. However, as we have noted, external testing is being mandated increasingly by state boards or legislatures as part of their efforts at educational reform.

There are two reasons the distinction between internal and external testing programs is important. First, external tests tend to have a greater mismatch with what is taught at the local level than do internal tests; for the simple reason that curriculum and teaching methods are determined, either explicitly or implicitly at the "grass-roots level" -- the LEA, the school and ultimately behind the classroom door. Second, external testing programs tend to have clearer consequences associated with them than do locally initiated and controlled tests. And as we argue in the next section the size of the stakes associated with testing programs is a key determinant of the educational impact.

4. High Stakes vs Low Stakes Testing Programs

A test whose results are seen - rightly or wrongly - by students teachers, administrators, parents, or the general public, to make important decisions that immediately and directly impact on them are what we shall term high-stakes tests. High-stakes student tests can be norm- or criterion-referenced, internal or external in origin. Examples include tests directly linked to such important decisions as: (1) graduation, promotion or placement of students; (2) the evaluation or rewarding of teachers or administrators; (3) the allocation of resources to schools or school districts; and(4) school or school system certification. In all of these examples, the perception of people that test results are linked to a high-stakes decision is in fact accurate. Policy makers have mandated that the results be used automatically to make such decisions.

However, there are other uses of test results that do not actually impact immediately and directly on students but nonetheless, are generally perceived by people as involving high-stakes. For example, SAT and ACT

results are of secondary importance in admission decisions in the vast bulk of colleges trying to fill vacant seats in the face of adverse demographics. Individuals and school systems, nevertheless, very often act on the perception that these college admissions tests are of crucial and singular importance, if not to individual students, at least to the public's peception of the quality of schooling. Thus, we find high schools are increasingly offering courses to prepare students to take these tests, and commercial coaching schools are doing a land office business.

In contrast to a high stakes test, a low stakes test is one which is perceived as having no important rewards or sanctions tied directly to test performance. Traditional school district standardized norm referenced testing programs, where results are reported to teachers, but there is no immediate, automatic decision linked to performance, are examples of this sort of testing. Teachers are free to ignore any results that they feel are discrepant from their own perceptions of students, and the results are not perceived by them as being used to evaluated their performance. This does not mean that test results from such programs do not affect teachers' perceptions of students, nor does it mean that student placement decisions are not related to test performance. The important distinction is that teachers, students, and parents do not perceive test performance as a direct vechicle of reward or sanction.

In short, it is clear that "high stakes" testing has the greatest impact on schooling, regardless of whether the stakes are associated with specific decisions made on the basis of the results, or with the perceived importance of the tests.

Arguments Concerning the Effects of Standardized Testing

If these are conditions which can affect the degree of impact of testing, what of the nature of the impact? Is it for good or ill? Over the last decade a wide variety of arguments have been advanced concerning the positive and negative effects of standardized testing. These arguments have been advanced in a variety of forums, including professional journals and meetings, the popular press, legislative debates, meetings of school governance agencies, and in both federal and state courts.

Without trying to document such myriad sources, it is useful to consider the nature of the arguments advanced regarding the positive and negative effects of standardized achievement testing. Among the most common affirmative arguments have been that such testing:

- helps focus instruction on skills (e.g. basic skills);
- motivates students;
- provides teachers with diagnostic information to improve instruction;
- identifies curriculum areas in need of improvement; and
- helps hold teachers and schools accountable for the learning of their charges.

Among the most common charges levelled against standardized achievement tests are that they:

- Are biased against certain kinds of students;
- Do not match what students have been taught;
- Constrain teachers' initiative and creativity in teaching;
- Promote teaching to the tests;
- Narrow the curriculum.

There are of course many variations on such arguments. And different arguments and their relative merits doubtless vary from place to place and program to program.

One point worth noting about such arguments, however, is that they suggest that standardized tests work their effects, for good or ill, on several different levels: by motivating individual students, by informing the instructional efforts of teachers, and by providing information which may guide improvement efforts at school, state or even national levels. This point is important, we think, because discussions of testing programs often seem to focus on the instrumental role of testing in which results may be used to help make specific decisions about students, programs or schools. Yet the arguments just cited suggest that indirect effects of testing are important, and we suspect, especially over time, may have a larger cumulative impact than direct effects related to specific decisions. What gets tested, after all, communicates to students, teachers, administrators and the community at large, what is valued in an educational endeavor in myriad ways. Thus, it is hardly surprising that with the increase in testing over the last decade or so, the skill of simply taking tests is emerging as a focus of concern, almost independent of what is ostensibly tested. A minor sign of the concern over test taking skills is that the phrase "Test taking skills" first emerged as a separate indexing category in volume 33 of the Education Index covering the July 1982 - June 1983 period; in the subsequent two volumes the number of items under this rubric has continued to grow.

III. PRINCIPLES ABOUT THE IMPACT OF TESTS

Perhaps the most convenient way to summarize the considerable literature on the impact of testing on education is to list major conclusions

from our review of this literature in the form of seven principles regarding the impact of testing.

Principle 1:

The more any quantitative social indicator is used for social decision-making, the more subject it will be to distort and corrupt the social processes it is intended to monitor.

This Principle comes directly from Donald Campbell's work on social indicators. It is not limited to testing per se; being much more general in scope, extending to any social indicator that is used to describe, make decisions about, or influence an important social process. This Principle reminds us that educational testing is a form of measurement subject to a social version of Heisenberg's uncertainty principle. Any measurement of the status an educational institution, no matter how well contrived, inevitably changes its status. And when the testing is used for important social decisions, the change tends to be both large and corrupting.

Principle 2

The power of tests and examinations to affect individuals, institutions, curriculum or instruction is a perceptual phenomenon: if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false -- the effect is produced by what individuals perceive to be the case.

Ben Bloom writing in the 68th NSSE Yearbook coined this second Principle. Its importance lies in the fact that when people perceive a phenomeon to be true, their actions are guided by the importance perceived to be associated with it. The greater the stakes perceived to be linked to test results the greater the impact on instruction and learning. A high-stakes test is one where the results are seen by students, teachers, administrators, parents, or the general public, to be used to make important decisions that

immediately and directly impact on individuals with whom they are concerned. Examples of high-stakes student tests include those perceived to be directly linked to such important decisions as: (1) graduation, promotion or placement of students; (2) the evaluation or rewarding of teachers or administrators; (3) the allocation of resources to schools or school districts; and(4) school or school system certification.

PRINCIPLE 3

If important decisions are presumed to be related to test results, then teachers will teach to the test.

This sort of accommodation by teachers to a high-stakes test is seen as having both positive and negative consequences. High-stakes tests, it is often argued, can focus instruction, giving students and teachers specific goals to attain. If the test is measuring basic skills, preparing students for the skills measured by the test could, proponents argue, serve as a powerful lever to improve basic skills. Unfortunately, however, the only sort of evidence generally available to bolster this proposition is that scores on high stakes tests do tend to increase over time. But standardized tests are indirect measures of the real skills of interest, and what repeated experience shows is that there are many, many ways of raising test scores without changing the levels of the skills the tests are intended to measure. People too often fail to distinguish between the skill and the indirect indicator of it.

If the test is specific to a specialized curriculum area, e.g. college preparatory physics, then the examination will eventually narrow instruction and learning, focusing only on those things measured by the tests. Indeed, this narrowing of the curriculum has been one of the enduring complaints leveled at external certification examinations used for the

important functions of certifying the successful completion of elementary or secondary education, and admission to third level education and/or certain jobs.

A review of the effects of such exams on the curriculum over many years, and in several countries, indicates that faced with a choice between objectives which are explicit in the curriculum or course outline, and a different set which are implicit in the certifying examination, students and teachers generally choose to focus on the latter. Spaulding's 1938 report pointed out that teachers in New York disregarded the objectives in local curriculum guides in favor of those tested in the Regents' examinations. Morris found the rigidity of the exams was the principal reason that the chemistry curriculum in Australia remained almost unchanged from 1891 to 1959. He concluded that the the proportion of instructional time spent on various aspects of the syllabus was "seldom higher than the predictive likelihood of its occurrence on the examination paper." Similar observations about the influence of the exams on the curriculum have been made in India, Japan, Ireland, and in England. Turner sums up the English experience: "One only has to look at the timetable of the typical comprehensive school to see that the curriculum consists almost entirely of subjects which can be taken in public examinations."

Why does this happen? First, there is tremendous social pressure on teachers to see to it that their students acquit themselves well on the certifying examinations. Second, the results of the examination are so important to students, teachers, and parents that their own self interest dictates that instructional time focus on test preparation.

As we have indicated, however, a high-stakes test can serve as a lever for the introduction of new curricular material. New curricula in physics, chemistry, and mathematics made an immediate impact in Irish schools when in the early 1960's they were prescribed and examined for the Leaving Certificate Examination. Primary teachers in Belgium accepted curricular reforms only when, in 1936, the external exams given at the end of primary school were modified to incorporate the ideas of the new curriculum. In New York State, curriculum specialists from the State Department of Education had little success in moving the emphasis in modern language teaching from grammar and translation to conversation and reading skills, until the corresponding changes had been incorporated in the content of the Regents' examination. Revisions of the College Entrance Examination Board (CEEB) math achievement tests to include modern math played an important part in the introduction of "modern" mathematics curricula in the 1960's.

Despite the ability of high stakes examinations to help introduce new material, the weight of examination precedent strongly influences not just what is taught and learned but also how. The question that educators must ask themselves is whether the positive aspects associated this phenomenon outweigh the disadvantages. The answer is a value judgment and depends on one's view of education, the learner, teaching, curriculum development and testing. Our view of the "driving" of instruction via external examinations is that in the long term, given the nature of most commonly used tests, the narrowing of instruction and learning associated with this phenomenon far outweigh any advantages.

Principle 4:

In every setting where a high-stakes test operates, a tradition of past exams has developed and it eventually comes to define the curriculum de facto.

Given Principle 3, the question still remains: "How do teachers cope with the pressure of the examination?" The answer is relatively simple. Teachers see the kind of intellectual activity required by previous test questions and prepare the students to meet these demands. Some have argued strongly that if the skills are well chosen, and if the tests truly measure them, then coaching is perfectly acceptable. This argument sounds reasonable, and in the short term, it may even work. But we need to take a longer view, because the argument ignores a fundamental fact of life: When the teacher's professional worth is estimated in terms of exam success, teachers have great incentive to produce gains; and they can do more simply by teaching test-taking strategies based on previous exam questions.

Further, the expectations and deep-seated primary agenda of students and their parents for exam success will further corrupt the process. The view that we can coach for the skills apart from the tradition of test questions is a staggeringly optimistic view of human nature which ignores the powerful pull of self-interest. It simply doesn't consider the long term effects of the examination sanctions.

An interesting aspect of Principle 4 is that if the examination is perceived as important enough, a commercial industry develops, outside the schools, to prepare students for it. In this country this phenomeon can be seen in the rise of commercial firms in virtually every major city selling coaching services to students for the SAT college admissions test. Another sign of the increasing prominence of test coaching in the United States is the fact that the phrase "test taking skills" first appeared as a separate indexing

category in volume 33 of the Education Index covering education literature for the July 1982 - June 1983 period. Most of the articles referenced under this new category dealt with improving admissions test scores through coaching provided by commercial or computer tutorial programs. In Japan it is common for parents to enroll their children in special extra-study schools known as *juku*. Beginning in the 19th century a whole industry of private coaching schools called "crammers" developed in Europe to prepare students, for a fee, for high stakes examinations. The important point about these coaching schools is not whether or not they are successful in preparing students for the exam, it is instead, that the public perceives them as helpful and willing to pay for their services.

Principle 5:

Teachers and students pay particular attention to the form of the questions on a high-stakes test, e.g. short answer, essay, multiple choice, and adjust their instruction accordingly.

The problem here is that the form of the test question can narrow instruction, study and learning to the detriment of other broader learning. Rentz recounts an example of this phenomenon which occurred as a result of the Georgia Regents' Testing Program, a program designed to assess minimum competencies in reading and writing on the part of college students in that state. The head of an English department lamented:

> Because we now are devoting our best efforts to getting the largest number of students past the essay exam ..., we are teaching to the exam, with an entire course, English III, given over to developing one type of essay writing, the writing of a five-- paragraph argumentative essay written under a time limit on a topic about which the author may or may not have knowledge, ideas, or personal opinions. Teaching this one useful writing skill has the beneficial effect of bringing large numbers of weak students to a minimal level of literacy, but at the same time, it devastates the content of the composition program that

> should be offering the better students challenges to produce writing of high quality. Because the Regents Test is primarily designed to establish a minimal level of literacy, our teaching to this test, which its importance forces us to do, tends to make the minimal acceptable competency the goal of our institution, a circumstance that guarantees mediocrity.

Principle 5 has profound implications for the curriculum specialists. Given our free enterprise system, publishers have begun to look at state mandated minimum competency or basic skills tests in order to design materials to better train pupils to take them. Children are apt therefore to find themselves spending more and more time filling out ditto answer sheets or work books. Deborah Meier, a successful principal of a public school in Manhattan, testified at the 1981 NIE sponsored hearings on Minimum Competency Testing (MCT) that in New York City, reading instruction has come to closely resemble the practice of taking reading tests. In reading class, students, using commercial materials, read dozens of little paragraphs about which they then answer multiple choice questions. Meier described the materials as evolving to resemble more and more the tests students will take in the spring. She went on to point out that when synonyms and antonyms were dropped from the New Yor City test of word meaning, teachers promptly dropped commercial material that stressed them. It is also interesting to note that in 1983, sales of ditto paper were way up nationally while sales of lined theme paper were down.

Principle 6:

<u>When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.</u>

Of all of the effects attributed to tests, this may well be the most damaging to education. It is illustrated in the following observation from a

nineteenth century British school inspector who observed first hand the negative effects of linking teacher salaries in England and Ireland to pupil examination results:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion as we tend to value the results of education for their measurableness, so we tend to undervalue and at last to ignore those results which are too intrinsically valuable to be measured.

Sixty years later, Ralph Tyler echoed the same message in the 62nd NSSE Yearbook when he warned readers that society conspires to treat marks in certifying examinations as the major end of secondary schooling, rather than as a useful but not infallible indicator of student achievement.

We see the importance society places on test scores to the exclusion of other indicators in such things as: the media attention to declines in SAT scores; reports that our schools score lower than those of other countries in math and science; the Education Department's wall chart that ranks states by their performance on the SAT or ACT; newspapers ranking school districts and/or schools within districts by their performance on standardized tests; the use of test results by real estate agents in selling homes; the money spent by parents on coaching schools for the SATs; the list could goes on and on.

Principle 7

A high-stakes test transfers control over the curriculum to the agency which sets or controls the exam.

The agency responsible for a high-stakes test assumes a great deal of power or control over what is taught, how it is taught, what is learned and how it is learned. This phenomenon is well understood in Europe where a

system of external certification examinations, controlled by the central government or by independent examination boards, operates at the secondary level. And while this shift in power is also understood in this country by policy makers who are mandating graduation and promotion tests, the implications of the shift from the local educational authority (LEA) to the state department of education (SEA) has not received sufficient attention and discussion. Further, since most state level testing programs are developed and validated for SEAs by outside contractors, it is important to realize that the state may be effectively delegating this very real power to a commercial company whose interest is primarily financial and secondarily educational.

Despite the negative consequences historically associated with Principles 1-7, the use of tests perceived as involving high-stakes, is growing in the United States. Policy makers are well aware of the high symbolic value of tests. By mandating a test they are seen to be making things happen. The test becomes a synecdoche for standards. The general public is gratified to find the tests restoring confidence in the schools. The numerical scores from high-stakes tests have an objective, scientific, almost magical persuasiveness about them, that the general public and policy makers are quick to accept. Curriculum specialists, teachers, and administrators therefore, increasingly are forced to deal with the consequences of mandated high-stakes tests. Until recently, these tests have generally involved basic skills or minimum competencies. However, recently various education reform reports have recommended a move to secondary school certification tests in individual subjects that closely resemble the British 'O' and 'A' examinations.

Haney and Madaus, paper for NAEP Study Group, 1986, p. 20.

In order to minimize the negative side effects associated with the impact of such tests, and to maximize possible benefits, it is important that professionals understand the value and limitations that historically have been associated with their use. They also need to understand how testing and the use of tests is changing in the United States: we are moving from LEA control to SEA control. Further, there are proposals that would alter the gathering and reporting of NAEP data to permit inter and intra state comparisons. The full philosophical, political and educational issues associated with such shifts in the use of tests need to be better understood and more fully debated.

Certainly, testing has an important function to play in American education. Test results when used in conjunction with other data about pupil performance, can help teachers to improve their instruction and to make educationally sound decisions. Test results can also provide important independent information to parents, the public and school administrators about the schools. Much of this information to which they are entitled. What we need right now is an awareness of the limitations and fallibility of tests. Test scores need to be regarded as one important element in decision making. The scores should be used by teachers and administrators in conjunction with other indicators of students' progress when making important decisions.

IV. NAEP as an Instrument for Informing Educational Policy

There are two basic ways in which test may affect educational policies. The first is the use of test information to <u>inform</u> policy makers about the current state of education. The second is the use of tests as <u>administrative devices</u> in the implementation of policy. In the former case test results are

used to describe the present state of education or some aspect of it, or in lobbying efforts for new programs or for reform proposals. The effects of this informational, descriptive use of test results on the educational process are indirect. This is in sharp contrast to the administrative use of test results where by results automatically trigger a direct reward or sanction being applied to an individual, or institution.

The Use of Tests Results to Inform Policy

The 1867 Act establishing the Department of Education recognized the need for gathering descriptive information about "the condition and progress of education in the several states and terrritories." Of course, at that time testing as we now know it did not exist. From the 1920's to the 1960's, standardized tests had little or nothing to do with state or federal policy. It was not until the early 1960's with the passge of the Elementary and Secondary Education Act of 1965, and the establishment of The National Assessment of Educational Progress (NAEP), that the Department began systematically to gather test data as part of its original mandate. Further, state departments of education have only recently begun to systematically collect test data to describe the status of education..

Their were several reasons for this shift. First, the concept of equality of educational opportunity evolved from a concern about equality of inputs, resources, and access to programs into a preoccupation with achieving equality of outcomes. As a result test scores began to be used as a primary indicator of educational outcomes. Second, advocates for minority groups began to point to the large discrepancies between the test scores of middle class students and their constituents to lobby successfully for compensatory funds for programs to reduce these disparities. Third, the large expenditures

in the 50's & 60's for curriculum development and compensatory programs led policy makers to ask for student test data as an indicator of the effectiveness of these programs. Fourth, as noted above NAEP was designed to provide a basis for public discussion and broader understanding of educational progress and problems of the nation.

More recently, the numerous educational reform reports have used test results, including SAT and NAEP data, to bring to the attention of the country what they conclude to be the mediocre state of American education, as well as to lobby for improvement programs to redress these weaknesses. Test data clearly form an important basis for the current negative descriptions of the status of American academic education. The question is how valid are these inferences? Is the academic performance of our students as poor as it is painted in the various reports? While there are weak spots - particularly at the secondary level with higher order skills - one could look at the same data and conclude that our schools are doing quite a creditable job; that declines and weak spots may be due in large part to non-school factors. In general the reports use test results in ways that accentuates the negative, eliminates the positive and leaves no room for Mr. In-Between.

An illuminating example of how these indicators are actually used to inform such reports is provided by the Twentieth Century Fund Report. It opens with the following gloomy assertion, "The nation's public schools are in trouble. By almost every measure..he performance of our schools falls far short of expectations". However, in a commissioned background paper, published as an appendix to the Report itself, Peterson examines all of the available indicators, including test scores, and concludes that, " Nothing in

these data permits the conclusion that educational institutions have deteriorated badly." It would seem that the Task Force did not take cognizance of its own commissioned paper.

Stedman and Smith in their excellent review of these reform proposals point out that they are quintessentially political documents. Testing evidence was used selectively to buttress arguments and evidencewas often ignored that might have lessened the impact of the message. Stedman andSmith examined critically the way test score indicators were interpreted and concluded not only was interpretation often sloppy, but also that we have little in the way of valid, longitudinal national indicators of the academic performance of students nationwide. NAEP, however, was cited as the single exception to the latter indictment.

V. The Future of NAEP

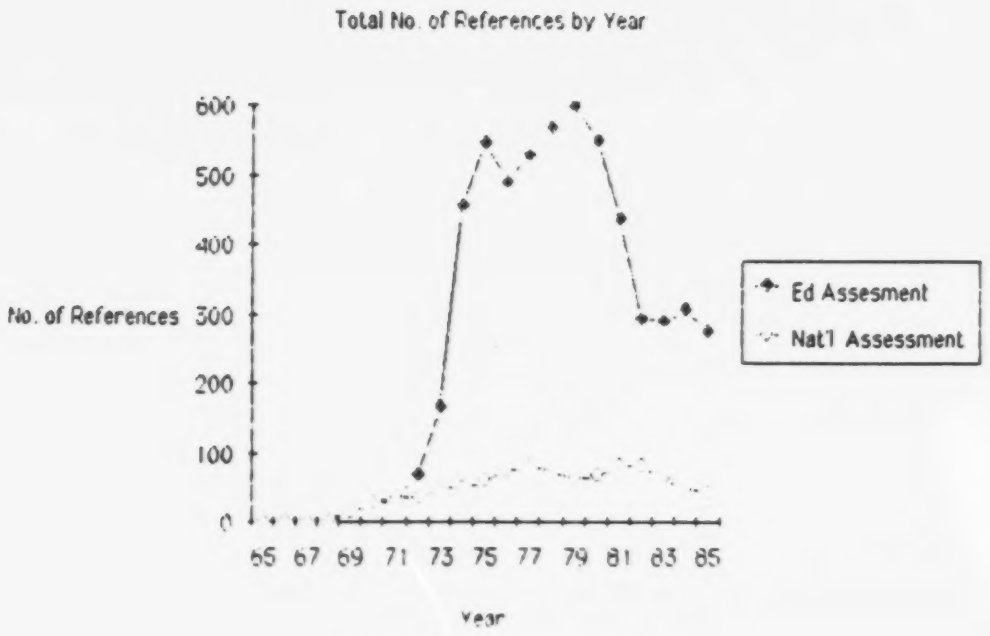
In considering the future of NAEP, we agree with Stedman and Smith and other observers of the currently available indicators of the nation's educational health; NAEP is the most valid source of <u>national</u> data on our students general achievement level over time. In light of our prveious discussion, we note too that NAEP has clearly been designed to <u>inform</u> educational policy-making, not drive it. This is in sharp contrast to many new testing programs implemented over the last decade at the state level which have been designed not just as means for informing policy, but instead as vehicles to <u>implement</u> educational policies. For instance, tests that are directly linked to decision-making about individual students (e.g. regarding grade to grade promotion, remediation or graduation), teachers (e.g. for teacher evaluation), schools or school systems (e.g. used allocate funding or other resources, or to rank publicly schools, school systems or

even states) all are, at least in some degree, instruments that not only inform policy, but also implement it.

Despite the relatively high praise that NAEP has received as a source of valid data on the status of our nation's students, it also has been subjected to a variety of criticisms. For example, the NAEP move from reporting reults at only the item level to reporting results aggregated across items, it was criticized for a serious failure to collect and analyze appropriate construct validity evidence (e.g. Haney, 1982). It is our understanding, however, that this problem has been -- or at least is being -- remedied via new latent trait scaling procedures designed by the Educational Testing Service since it has taken over the administration of NAEP from the Education Commission of the States .

A more common criticism of NAEP by almost every review of it over the last two decades has been that its results have not been terribly useful (e.g. Greenbaum, Garet and Solomon, 1977, Hazlett, 1974, , GAO, 197, Wirtz and Lapointe , Haney 1982). In order to provide a rough check on the utility of NAEP over time we performed a search of the ERIC data base to see the number of occurrences of the expression "national assessment" by year. To provide some perspective we also ascertained the total number of citations in the ERIC data base by year, and as a points of contrast with the search for "national assessment" items, the number of occurrences, also by year of items containing the phrase "educational assessment." Details of how we conducted this search and its limitations are described in Appendix 2, and the results are depicted in Figure 2.

Our findings suggest two major points. One is that in the citations from around 1970 (specifically 1969, 1970, and 1971), the number of occurences of the phrase "national asessment" exceeded the number of


Total No. of References by Year
600
500
400
300
200
100
0
No. of References
65 67 69 71 73 75 77 79 81 83 85
Year
Ed Assesment
Nat'l Assessment

occurrences of the phrase "educational assessment." However, beginnning in 1973 the number of times the phrase "educational asessment" is found in the ERIC databases increases dramatically. This is evidence, we suspect of one of the indirect affects of NAEP, namely that of promoting broader attention in the educational research community to the idea of educational assessment. More concretely, of course, the NAEP model also contributed in the 1970s to the founding of a number of statewide assessment programs. (Given the scale of the vertical axis of Figure 2, it is hard to discern the pattern of citations, for the pre-1973 years, but readers interested in details may consult the data table in the appendix.)

More recently, in the mid-1980s, the number of entries under both terms "educational assessment" and "national assessment" have decreased sharply, roughly one-third to one-half from what they were at their peaks (in 1982 the peak number for "national assessment" was 83, and for "educational assessment" the peak occurred in 1979 with 598). In part these changes may simply represent the trend of enthusiasms and jargon in the broad field of educational reasearch and testing. For it seems apparent that the enthusiasm for "assessment" in the 1970s gave way to interest in "competency testing" and "basic skills testing" in the late 1970s and early 1980s.

Nevertheless, the figures represented in Figure 2 give cause for concern about the continuing utility of NAEP. And it is against this background that we wish to consider ideas about the future of NAEP. Specifically, we wish to consider the idea of extending NAEP to be used in statewide testing programs, and the possibility of NAEP assessment exercises being used not just in sampling assessments but also in programs in which

whole populations of students are tested (e.g. within a state or school district).

The notion of using NAEP exercises (or items) not just in national assessments based on stratefied random samples, but also in statewide testing programs in which all students at a particular grade might be tested makes eminent economic sense. It would allow the considerable costs of development of NAEP exercises to be amortized over much larger populations. However, what we wish to warn against are three dangers associated with the possibility of using NAEP in statewide assessments (particularly where whole populations of students rather than samples might be tested).

The dangers we wish to warn against are : 1) If NAEP becomes an instrument of educational policy it likely would have distorting effects on local programs; 2) If NAEP becomes an instrument of policy, it would threaten the validity of NAEP findings, and 3) Such wider instrumental use of NAEP might threaten some of the indirect benefits which we think have been associated with NAEP in its first 20 years.

1. NAEP as an instrument of educational policy would have distorting effects. Here our worry is that in the efforts to make NAEP more useful, if it develops into a program which would allow specific decisions to be made on the basis of NAEP results (or even state by state comparisons), it would change from what it has been over the last 20 years, mainly a source of high quality information for informing educational policy, into an instrument of policy.

In short our concern is that to the extent that NAEP becomes an administrative device for implementing educational policy, the negative effects described in the seven principles above would come into play. We

should hasten to add that we know of no specific plans by anyone to use NAEP for making specific decisions about students (such as grade promotion or graduation). However, as we have argued, even comparisons such as state or district rankings on the basis of test results can lead to distortions.

2. NAEP as instrument of policy, might have less validity. Our second concern is that efforts to improve the utility of NAEP migth actually threaten the validity of assessment data derived from it. Here, we note that if there has been one feature of NAEP universally praised, it is the high quality of information derived from NAEP assessments (with the notable exception, prior to the ETS administration of NAEP, of the failure to gather construct vality evidence supporting the reporting of data on groups of NAEP exercises -- but as we said, we think that this problem has been largely solved with new ETS scaling procedures). However, if efforts to make NAEP more useful have the effect of making NAEP an administrative instrument of educational policy, this would likely not only have distorting effects on educational practices, but also would threaten the validity of assessment data derived from NAEP. Another way of communicating this concern is to note the distinction between obstrusive and unobtrusive measurment. One of the real virtues of unobyrusive measurement is that because it is unobtrusive it tends to have higher measurement validity than obtrusive measurement. Unobstrusive measurement need not be unuseful, of course, but it is our view that educational testing and assessment works best and has fewer untoward consequences when their effects are mediated through the informed judgments of educational policymakers and practitioners.

3. Indirect Benefits

As we have already noted, one of the relatively early indirect benefits of NAEP seems to have been that in the mid-1970s, it helped promote the development of broader interest in educational assessment and specifically several statewide assessment programs. However, in the late 1970s as we also already noted, many state programs moved in the direction of "minimum competency" or "basic skills" testing. In this regard we feel that one of the real indirect benefits of NAEP is that it has provided a model of broader assessment, broader both with regard to the range of knowledge and skills tested, and with regard to the methods of assessment used.

In this regard and in thinking about the possible indirect effects of standardized testing on education in general and on curriculum specifically, two general points seem apparent. One is that though most standardized tests may focus on reading and math skills, the general aims of education as represented in curriculum are considerably broader. Take for example the "New Basics" advocated by the National Commission on Excellence in Education (1983). This group, though only one of many groups recently advocating reforms of education, was perhaps the most prominent. It recommended that:

> State and local high school graduation requirements be strengthened and that at a minimum, all students seeking a high school diploma be required to lay the foundations in the Five New Basics by taking the following curriculum during the 4 years of high school: (a) 4 years of English; (b) 3 years of math; (c) 3 years of science; (d) 3 years of social studies; and (e) one-half year of computer science. For the college-bound, 2 years of foreign language are strongly recommended in addition to those taken earlier. (National Commission on Excellence in Education, 1983, p. 24.)

This is of course only one of many recent recent reform recommendations for improving education in the United States. But a striking point about this recommendation, in contrast to the coverage of most "standardized" tests, is that three of the five "new basics" are badly, if not completely, neglected in most statewide testing programs. This has not been the case with NAEP.

A second general feature of almost all large scale testing programs, with NAEP again being perhaps the most notable exception, is that they employ the multiple-choice format almost exclusively. This format is in striking contrast to the calls in many of the recent education reform reports that students should be taught not just to solve pre-set problems, but to identify and work out solutions to problems they find for themselves. In this regard we are reminded of Norman Frederiksen's article suggesting that the predominance of the multiple-choice format represents "the real test bias". Frederiksen reviews evidence showing that particuliarly for higher order problem solving skills, invioving solving open-ended or ill-structured problems, multiple-choice items may not be good measures of the broad domain of skills and abilities that we might hope are taught in schools. Thus, the preponderance of tests employing the multiple-choice format, which surely has contributed to recent interest in "test-taking skills," may be communicating exactly the wrong message about what it is that we want our schools to be teaching. In this regard it is notable that Frederiksen closed his article on the real test bias with a message remarkably similar to the one from Ralph Tyler fifteen year earlier, quoted at the start of this article. As Frederiksen put it:

> The "real test bias" in my title has to do with the influence of tests on teaching and learning. Efficient tests drive out less efficient tests, leaving many important abilities untested anbd untaught. An important task for those involved in testing is to

> develop instruments that will better reflect the whole domain of educational goals and to find ways to use them in improving the educational process. (p.19)

In the future, we therefore hope that there will be maintained the past tradition of NAEP, of employing diverse methods of assessment and helping to do what both Tyler and Frederiksen advocate -- namely making sure that our assessments span as broad a range of the curriculums and goals of our schools as possible.

Appendix 1

The Education Index, published by the H. W. Wilson Company since the 1932 is an author and subject index to educational material in the English language. Though primarily a periodical index it also covers proceedings, yearbooks, monographs and material printed by the U.S. Government. In terms of periodical coverage, the index covers both professional and more popular journals. Actual selection of periodicals for indexing is accomplished by subscriber vote, at least since 1970. In voting , subscribers are asked to place primary emphasis on the reference value of periodicals under consideration.

Figure 1 was constructed simply by measuring the number of column inches devoted to listings concerning testing and curriculum in every volume of the Index from1932 (volume 1, covered material published from January 1929 - June 1932) through 1965 (volume 33: July 1964 - June 1965). Over these volumes there were ocassionally some changes in the index rubrics concerning testing and curriculum. However the primary rubrics considered as to pertaining to the two topics of interest were:

Testing:
- Tests and Scales
- Testing programs (introduced in vol. 6)
- Tests of general educational development (vol. 23)
- Testing instruments (vol. 25)

Curriculum:
- Curriculum
- Curriculum Making
- Curriculum laboratories (introduced in vol. 4)
- Curriculum satisfaction (vol. 8)
- Curriculum selection (vol. 8)
- Curriculum development (replacing Curriculum Making in vol. 11)
- Curriculum studies (vol. 16)

It should be noted too that only since 1964 (volume 14 covering 7/63-7/64) has the Index been produced annually. Before that time it was issued on either biennial or triennial bases. Thus for years prior to 1964 we have shown the average annual numbers of column inches listed under the relevant testing and curriculum rubrics. Finally it is worth mentioning that over all 33 volumes of the Index, the type size has remained the same and that 1 column inch, including headings and subheadings, amounts on average to roughly 2 to 3 references.

References

Anastasi, A. (Ed.) 1966. Testing problems in perspective. Washington, D.C.: American Council on Education.

Congressional Budget Office, 1986. Trends in Educational Achievement. Washington, D.C.: author.

Education Index. 1932 -1985. Voulmes 1 - 35. New York: H. W. Wilson.

Frederiksen, N. 1981. The real test bias. (ETS Research report 81-40). Princeton, NJ: Educational Testing Service.

Pipho, C. 1985. Tracking the reforms: Part 5:Testing -- Can it measure the success of the reorm movement? Education Week, May 22, 1985, p. 19.

National Commission on Excellence in Education. A Nation at Risk: The Imperative for Educational Reform. Washington, D.C.: U.S. Department of Education.

Tyler, R."What Testing Does to Teachers and Students," 1959, In Anastasi, 1966, pp. 46-52. .

ERIC REFERENCES CONCERNING ASSESSMENT

ERIC Citations. Total. Regarding Ed'l Ass't. Nat'l Ass't. By Year

Year	Total Ref's	Ed'l Ass't	Nat'l Ass't	Ratio #
65	2959	1	0	0.00
66	4856	5	1	0.20
67	7115	6	0	0.00
68	9824	9	2	0.22
69	25317	15	16	1.07
70	28954	26	30	1.15
71	32499	36	37	1.03
72	34063	68	35	0.51
73	35208	164	50	0.30
74	35274	453	50	0.11
75	38168	545	59	0.11
76	37417	489	76	0.16
77	36914	528	82	0.16
78	38471	564	75	0.13
79	35984	598	64	0.11
80	34591	548	66	0.12
81	31516	436	80	0.18
82	30661	293	83	0.28
83	30506	290	64	0.22
84	29591	306	45	0.15
85	24275	275	44	0.16